Executive Steering Committee
For A.C.E. Policy II
(ESCAP II)
Report 19
September 27, 2001

ESCAP II

Analysis of Non-Matches and Erroneous Enumerations Using Logistic Regression

Michael Beaghen
Roxanne Feldpausch
Rosemary Byrne
Decennial Statistical
Studies Division

U S C E N S U S B U R E A U
Helping You Make Informed Decisions

CONTENTS

LIST OF TABLES

EXECUTIVE SUMMARY

We conducted this evaluation to aid in the understanding of Census erroneous enumerations and Census omissions in the 2000 Accuracy and Coverage Evaluation Several univariate analyses of Accuracy and Coverage Evaluation variables have already been done. The multivariate analyses we conducted complemented these analyses. We built two logistic regression models; one used E-sample data to relate erroneous enumerations to demographic and geographic variables; the second used the P-sample data to relate census capture to demographic and geographic variables.

What did the model tell us about the efficacy of the Accuracy and Coverage Evaluation poststratification?

Although the modeling was not designed to evaluate the efficacy of the poststratification, **it provided some evidence that the Accuracy and Coverage Evaluation poststratification was appropriate.** Nothing we found suggested the poststratification was inappropriate.

- The poststratification was by Age/Sex, Racial/Ethnic Domain, Tenure, Region, Metropolitan Statistical Area/Type of Enumeration Area, and Mail Return Rate. The model showed the predictive value of all these variables except Mail Return Rate, which was not included in our models.
- Generally, the people most likely to be correctly enumerated were also most likely to be captured. That is, groups who had greater odds of correct enumeration also had greater odds of capture. This was true for the groups defined by the variables Relationship, Size of Household, Type of Basic Street Address/Number of Units at Basic Street Address, Race/Ethnic Domain, Tenure, Region, and Metropolitan Statistical Area/Type of Enumeration Area. However, Age/Sex did not fit this pattern.

Did the multivariate model find the same variables were important that the univariate analyses found important?

Yes. The multivariate model found the same variables were important that the univariate analyses did. The E-sample model showed that the following variables were associated with the probability of a person being erroneously enumerated:

- Relationship was important in its predictive value. The reference person and their spouse had almost twice the odds of correct enumeration as other household members.
- Size of Household: people in single person households had smaller odds of being correctly enumerated than those in multi-person households.
- Number of Units at Basic Street Address: people in single units had greater odds of being correctly enumerated.
- Age/Sex and the subsequent variables were all significant but weaker effects. Younger people had greater odds of being correctly enumerated.
- Tenure: owners had greater odds of being correctly enumerated.
- Racial/Ethnic Domain: Whites, Hispanics and Asians had greater odds of being correctly enumerated than Blacks.

- Region: people in the Midwest forms had greater odds of being correctly enumerated.
- Type of Enumeration Area (TEA): people enumerated in mailout/mailback had greater odds of being correctly enumerated.
- The variable Form Type had no effect on odds of being correctly enumerated.

The P-sample model showed that the following variables were all associated with the probability of an Accuracy and Coverage Evaluation person being a non-match:

- Relationship stood out in its predictive value. The reference person and their spouses had about 1.25 times greater odds of capture than their children, and about twice the odds of capture as non-relatives or other relatives.
- Age/Sex was the next most important variable, with older people and females having a greater odds of capture. In particular, females 50 years old or older had about 1.8 times as great odds of capture as 18-29 year old men.
- Size of Household was also relatively important. People in households of size two to six had greater odds of capture than people in a single person household. In households with seven or more people reference people had a high odds of capture. However, the non-reference people in these households had the lowest odds of capture.
- Tenure and the subsequent variables were weaker in their effects, though owners had greater odds of capture.
- Region: people in the Midwest had greater odds of capture.
- Racial/Ethnic Domain: Whites were more likely to be counted than other groups.
- Type of Basic Address: people in single unit structures had greater odds of capture.
- Type of Enumeration Area: people in mailout/mailback type of enumerations areas had greater odds of capture.

What were the advantages of the multivariate model?

The multivariate model allowed us to account for correlations between variables and to model interactions. When we controlled for the effects of correlations between variables we gained the following insights:

- Effects often seemed stronger in the univariate analysis than in the model because the univariate analysis did not take into account the correlation between the variables.
- The relationships between the variables Tenure, Number of Units at Basic Street Address, and Racial/Ethnic Domain with the odds of correct enumeration were shown by the E-sample model to be weaker than they appeared in the univariate analysis.
- The relationships between the variables Age/Sex, Tenure, Type of Basic Address, Racial/Ethnic Domain, and Region with odds of capture were shown by the P-sample model to be weaker than they appeared in the univariate analysis.
- When one controlled for Household Size, spouses were no more likely to be correctly enumerated or captured than the reference person. This finding contrasted with the univariate analysis, which suggested spouses were more likely to be correctly enumerated and counted. In the univariate analysis reference people were confounded with single person households, which were more likely to be erroneously enumerated and not captured.

The model also identified the following interactions:

- Non-reference people had much smaller odds of capture in households with seven or more people. This was because the census form only accommodated person information for six respondents, and the additional household members were often not data-defined.
- While the people in the Midwest were more likely to be captured in general, Blacks and Hispanics were no better counted there than in other regions.
- People in small multi-units (2-9 units) were more likely to be erroneously enumerated in the Northeast than in other regions.
- People in multi-units were better captured in the West than in other regions.
- Adult children were more likely to be erroneously enumerated than young children.
- Black males age 18-49 were slightly more likely to be erroneously enumerated than one would otherwise predict.

What implications do these results have on the adjustment decision?

While the results of the logistic regression modeling did not bear directly on the question, the fact that **things were as we expected** reassures us about the quality of the Accuracy and Coverage Evaluation.

1. BACKGROUND

This evaluation analyzed data from the Accuracy and Coverage Evaluation (A.C.E.) to gain insights into Census erroneous enumerations and Census omissions.

1.1 The Accuracy and Coverage Evaluation (A.C.E.) Survey

The A.C.E. measured the accuracy of the Census 2000 (Childers 2001). The A.C.E. consisted of two samples: a sample of people, the P sample, and a sample of census enumerations, the E sample. The P sample was obtained by conducting an independent enumeration of people in sampled clusters of census blocks. There were 721,734 P-sample people in 11,303 block clusters nationwide (Puerto Rico not included). The P sample measured the census coverage or miss rate. The E sample consisted of those census enumerations in the A.C.E. sampled clusters. There were 712,900 E-sample people. The E sample measured the census erroneous enumeration rate. The miss rate and the correct enumeration rate, taken together with the census count of enumerations eligible to be selected in the E sample, produced a dual system population estimate.

Central to the A.C.E. was a matching operation that compared the P-sample records to E-sample records and a field followup to resolve differences.

- People found in both the P sample and the census were called a match.
- People found only in the P sample and confirmed to be Census Day residents were called non-matches and represented census misses or failures to capture.
- E-sample people who matched to P-sample people who were residents were correct enumerations.
- E-sample people who did not match were searched for duplication.
- E-sample people not matched to a person in the P sample and not duplicated were followed up to determine whether they were correct enumerations or erroneous enumerations. If the non-matched census person was found to have lived in the A.C.E. sample cluster on Census Day, he or she was a correct enumeration. If the census person was found not to have lived in the A.C.E. sample cluster on Census Day, he or she was an erroneous enumeration.

Erroneous enumerations were as follows:

- Duplicated people
- People who were not living as residents in the A.C.E. block cluster on Census Day
- People who were beyond the search area and thus were geocoding errors
- Person records that did not correspond to people in the search area
- Census people with less data than a good name and two characteristics were counted as erroneously enumerated and were likewise not eligible to be matched to the A.C.E. (The dual system estimates of population were unbiased so long as the census population to which the P-sample people could match was the population of census people meeting the A.C.E. definition of a correct enumeration.)

1.2 The Purpose of this Evaluation

The purpose of this paper was to build two logistic regression models; one to relate census misses as identified by P-sample non-matches to variables such as person demographic characteristics, housing unit characteristics and census enumeration methods, and a second to relate erroneous enumerations to similar variables. While univariate descriptive statistics were illuminating, they did not address the question of the relationship of one variable to the response in the context of other variables. The multivariate models avoided this limitation and thus complemented the univariate studies being done. Logistic regression was an appropriate multivariate method since the responses in both models were binary; for example, for the P-sample model a person was either captured or missed by the census, for the E-sample model a person was a correct enumeration or an erroneous enumeration.

1.3 Logistic Regression Modeling

A logistic regression model takes the following form; the response is defined as a success (that is, correct enumeration or census capture) or failure (that is, erroneous enumeration or census miss). Logistic regression then models the natural logarithm of the odds of a success. The odds of success are related to the probability of a success by $p_i/(1 ! p_i)$ = odds, where p_i is defined as the probability of a success for the i^{th} individual. The k parameter logistic regression model is:

$$\log(p_i/(1 ! p_i))=\beta_0+\beta_1 x_{i1}+\text{þ}+\beta_k x_{ik}.$$

2. METHODS

The variables in the model included demographic ones such as the variables in the A.C.E. poststratification (Griffin 2000). All variables were categorical. The variables used for both models were Tenure, Age/Sex, Size of Household, Racial/Ethnic Domain, Type of Enumeration Area and Region. For the E sample we used the Number of Units at the Basic Street Address, which was classified into single unit, 2-9 units, or 10 or more units, consistent with the A.C.E. housing unit dual system estimation poststratification. For A.C.E. modeling we chose type of basic address over the number of housing units at a basic street address. These variables were nearly collinear, though type of basic address had slightly more explanatory power. For details on the source data files and any recoding of variables, see the Technical Documentation in the Appendix.

We handled census people with unresolved codes by considering a person to be a correct enumeration if the imputed probability of correct enumeration was 0.5 or greater and an erroneous enumeration if it was less than 0.5. Similarly, a P-sample person with an imputed probability of match greater than or equal to 0.5 was counted as a match and one with a probability less than 0.5 was considered a nonmatch. P-sample people with unresolved residence status were handled by incorporating the probability of residence into the person's weight.

The odds ratios for univariate models were also calculated and shown for comparison. To obtain the odds ratios for the univariate analysis we estimated a separate logistic regression model for each variable, where the only regressor in the model was that variable.

For both models the data was weighted by the sampling weights and the estimates and hypothesis tests reflected the complex sampling structure of the data. See the Technical Documentation in the Appendix for details on the estimation of the models.

3. LIMITS

Note that this study was observational rather than experimental. The characteristics used as regressors in the model were not controlled by the researcher but rather were random variables in themselves. Consequently the modeling was not predictive but descriptive and causal relationships between variables could not be inferred. Furthermore, the hypothesis tests used to determine which variables to include in the model were conditioned on the regressors and were therefore not strictly correct.

There were several limitations in our method for identifying interactions. For details on this methodology see the Technical Documentation in the Appendix. First, our methods were vulnerable to finding spurious interactions statistically significant because we screened many interactions. At the same time, our methods may have failed to detect statistically significant interactions of small magnitude. While we were confident we found the most important first order interaction terms, we may have missed less important though statistically significant ones. These limitations were not serious for this research because its purpose was descriptive not predictive. Further, with one exception the importance of interactions in the model was not large in any case. This was seen when we examine the Wald-Chi Square test statistics of the interaction terms.

Lastly, note that the P-sample analysis did not reflect the number of inmovers. The number of inmovers was reflected in the A.C.E. production estimation of coverage.

4. RESULTS

In this section we present the results of using logistic regression models to analyze the A.C.E.'s E-sample and P-sample data. Note that in the following discussions we use the word variable to describe both categorical variables such as Racial/Ethnic Domain, which has seven categories, and the zero-one indicator terms that parameterize the categorical variables.

4.1 Interpreting the Results of Modeling the E Sample

The A.C.E. E sample was designed to measure census erroneous enumerations. We used a logistic regression model to relate demographic and other variables of interest to the probability of an erroneous enumeration. Also in this section we explain how the logistic regression modeling was interpreted.

4.1.1 How to interpret the logistic regression model

Rather than examining the parameter weights themselves it was easier to interpret the odds ratios associated with an increase of one unit for each variable, which were directly related to the parameter weights (Hosmer, Lemeshow 1989). Because the standard errors of the odds ratios are

not symmetric about the estimate, 95 percent confidence intervals are shown instead. Since each of the variables had a value of zero or one depending on the level, the interpretation of the odds ratio is straightforward. The odds ratio refers to the ratio of the odds of a response with a value of one to the odds of a response with a value of zero. The absolute value of the odds ratio only makes sense in comparison to reference levels. However, the odds ratios between levels of the same variable can be compared directly. As an illustration consider Relationship. As shown in Table 1, six of the Relationship categories were indicated by variables which have estimates, whereas one category, reference person, was the reference level and was set to 1.0. Now, the odds ratio of 0.442 associated with Sibling implied that a person who was a sibling had only about 44 percent the odds of being correctly enumerated as the reference person, all other variables held constant. It also implied that a person who was a parent had about 28 percent greater odds (0.565/0.442) being correctly enumerated than did a sibling.

The Wald Chi-square test statistic gave an indication of the relative importance of a variable. Relationship, with a Chi-square of 873.6, was a dominant variable. Thus Relationship, along with Size of Household and number of Units at the Basic Street Address obtained most of the discrimination of the model.

Note that all variables and interaction terms were significant at the 0.1 alpha level.

Table 1. Results of E-Sample Modeling

Variables	Wald Chi-Square	Odds Ratio	95% Confidence Interval (Model)	Univariate Odds Ratio
Relationship	873.6			
Reference Person		1.000	reference level	1.000
Spouse		0.880	(0.848, 0.912)	1.302
Child		0.735	(0.658, 0.821)	1.008
Sibling		0.442	(0.399, 0.490)	0.463
Non-Relative		0.460	(0.433, 0.489)	0.491
Parent		0.565	(0.499, 0.639)	0.691
Other Relative		0.465	(0.447, 0.548)	0.679
Size of Household	229.5			
2 - 6 People		1.000	reference level	1.000
Single Person		0.626	(0.589, 0.666)	0.630
7 or more People		1.144	(1.015, 1.289)	0.908
Units at Basic Street Address	185.8			
Single Unit		1.000	reference level	1.000
2-9 Units		0.525	(0.478, 0.578)	0.365
10+ Units		0.645	(0.584, 0.713)	0.479

Variables	Wald Chi-Square	Odds Ratio	95% Confidence Interval (Model)	Univariate Odds Ratio
Age/Sex	149.9			
50+ Female		1.000	reference level	1.000
0-17		1.302	(1.156, 1.467)	1.100
18-29 Male		0.956	(0.883, 1.034)	0.609
18-29 Female		1.023	(1.022, 0.946)	0.688
30-49 Male		0.979	(0.918, 1.045)	1.133
30-49 Female		1.145	(1.077, 1.217)	0.933
50+ Male		0.834	(0.795, 0.875)	0.959
Tenure	102.8			
Owner		1.000	reference level	1.000
Non-Owner		0.706	(0.660, 0.755)	0.476
Racial/Ethnic Domain	102.0			
Non-Hispanic White		1.000	reference level	1.000
American Indian on Reservation		1.087	(0.897, 1.318)	0.964
American Indian off Reservation		0.773	(0.631, 0.947)	0.649
Hispanic		0.960	(0.889, 1.037)	0.725
Non-Hispanic Black		0.703	(0.654, 0.755)	0.542
Native Hawaiians or Pacific Islanders		0.775	(0.775, 0.555)	0.578
Asian		0.895	(0.796, 1.006)	0.752
Region	30.6			
West		1.000	reference level	1.000
Midwest		1.151	(1.050, 1.263)	1.262
Northeast		1.165	(1.045, 1.298)	0.957
South		0.918	(0.842, 1.001)	0.946
Type of Enumeration Area	27.6			
Large MSA, Mailout/Mailback		1.000	reference level	1.000
Medium MSA, Mailout/Mailback		1.030	(0.935, 1.133)	1.159
Small MSA and Non-MSA MO/MB		1.017	(0.914, 1.133)	1.197
All other TEAs		0.833	(0.748, 0.930)	1.128
Interactions				
Age 18+, Relationship Child	40.6	0.668	(0.590, 0.756)	N.A.
Northeast, BSA 2-9 Units	27.0	0.693	(0.604, 0.796)	N.A.
Males 18-49, Black	3.5	0.930	(0.861, 1.003)	N.A.

4.1.2 Interpreting the variables

In this section we go over each variable and their interactions.

- The most important variable was **Relationship**. The reference person and their spouses had about 1.25 times odds of correct enumeration than their children and nearly twice the odds of being correctly enumerated than other household members.
- **Size of Household**: people in single person households had smaller odds of correct enumeration by a factor of about 0.6.
- **Units at Basic Street Address** showed that people in single units were more likely than people in multi-units to be correctly enumerated. The odds ratios were 0.525 and 0.645 for people at basic street addresses with 2-9 units and 10 or more units. Note that in the Northeast the odds of correct enumeration for people in structures with 2-9 units were even smaller as they were multiplied by a factor of 0.693 (see the Northeast, BSA 2-9 Units interaction).
- The **Age/Sex** variable showed that 0-17 year olds had greater odds (1.302) of correct enumeration than 50 years and older people. Sex played only a small role in the probability of correct enumeration.
- There was an interaction between **Relationship** and **Age/Sex**. People whose relationship was Child had smaller odds of being correctly enumerated by a factor of 0.688 if they were also 18+ years old.
- **Racial/Ethnic Domain**: Blacks were a little less likely to be correctly enumerated than Whites. American Indians on Reservation were more likely to be correctly enumerated.
- The interaction term **Males 18-49, Black** had an odds ratio of 0.928. Thus Black men in this age group had only about 93 percent the odds of being a correct enumeration than we would have predicted based only on the two associated main effects, that they were Black, and that they were males 18 to 49.
- **Region**: people in the Midwest were more likely to be correctly enumerated. The Northeast had an odds ratio close to that of Midwest. However, when the Basic Street Address was 2-9 units people in the Northeast had an odds ratio decreased by a factor of 0.693; see the **Northeast, BSA 2-9 Units** interaction.
- **Tenure**: Non-owners had a slightly lower odds of correct enumeration than did owners.
- **Type of Enumeration Area**: people in non-mailout/mailback types of enumeration areas were less likely to be correctly enumerated with an odds ratio of 0.833.
- Note that the variable **Form Type** was found to be insignificant. It was taken out of the model and thus it is not shown in Table 1.

Comparing the model odds ratios with the univariate odds ratios one notices that the effects were often stronger in the univariate analyses. (When comparing two odds ratios keep in mind that a larger value of the odds ratio indicates a stronger effect if both odds ratios are greater than one, but a weaker effect if both ratios are less than one). The odds ratio for Tenure in the model, 0.706, indicated a smaller effect than did the odds ratio estimated in the univariate analysis, 0.476. For the Units at Basic Street Address variable, the odds ratios for the levels of multi-unit versus single unit given by the model (0.525 and 0.645) were also more modest than that estimated in the univariate

model (0.365 and 0.479). Similar held true for the level Black of the variable Racial/Ethnic Domain (0.703 versus 0.542). This exaggeration of the effect observed in the univariate model was seen because the univariate model did not take into account the correlations between regressor variables. In this case, Tenure, Units at Basic Street Address and Racial/Ethnic Domain were all correlated. Note that since interactions were not examined in univariate models there was no applicable (N.A.) univariate odds ratio for comparison for interaction terms.

4.2 Interpreting the Results of Modeling the P sample

When modeling the P sample we were modeling the probability of census capture instead of the probability of correct enumeration. Otherwise the interpretation of the P-sample model logistic regression results (given in Table 2) was analogous to that of the E-sample model.

Note that all variables were significant at the 0.1 alpha level.

Table 2. Results of P-Sample Modeling

Variables	Wald Chi-Square	Odds Ratio	95% Confidence Interval (Model)	Univariate Odds Ratio
Relationship	1521.5			
Reference person		1.000	reference level	1.000
Spouse		1.063	(1.034, 1.092)	1.412
Child		0.827	(0.788, 0.868)	0.922
Sibling		0.532	(0.486, 0.581)	0.438
Parent		0.701	(0.639, 0.769)	0.816
Other relative		0.392	(0.366, 0.420)	0.345
Non-relative		0.423	(0.401, 0.446)	0.364
Missing		0.607	(0.537, 0.686)	0.605
Age/Sex	445.6			
50+ female		1.000	reference level	1.000
0-17		0.829	(0.771, 0.891)	0.625
18-29 Male		0.585	(0.548, 0.623)	0.477
18-29 Female		0.681	(0.638, 0.727)	0.391
30-49 Male		0.687	(0.650, 0.725)	0.807
30-49 Female		0.815	(0.771, 0.861)	0.638
50+ Male		0.823	(0.787, 0.861)	0.897
Size of Household	316.7			
2 - 6 People		1.000	reference level	2.630
Single Person		0.629	(0.597, 0.662)	1.761
7 or more People		1.103	(0.942, 1.292)	1.000

Variables	Wald Chi-Square	Odds Ratio	95% Confidence Interval (Model)	Univariate Odds Ratio
	277.1			
Tenure				
Owner		1.000	reference level	1.000
Renter		0.617	(0.583, 0.653)	0.431
Type of Basic Address	158.6			
Single		1.000	reference level	1.000
Multi-unit		0.671	(0.617, 0.729)	0.450
Trailer not in park		0.601	(0.516, 0.699)	0.475
Trailer in park		0.473	(0.375, 0.597)	0.413
Racial/Ethnic Domain	110.0			
White		1.000	reference level	1.000
American Indian on reservation		0.770	(0.618, 0.959)	0.536
American Indian off reservation		0.753	(0.611, 0.927)	0.443
Hispanic		0.828	(0.767, 0.894)	0.518
Black		0.698	(0.650, 0.750)	0.488
Native Hawaiian or Pacific Islander		0.566	(0.383, 0.837)	0.399
Asian		0.872	(0.771, 0.987)	0.708
Region	89.8			
West		1.000	reference level	1.000
Northeast		1.093	(0.974, 1.225)	1.041
Midwest		1.472	(1.315, 1.649)	1.468
South		0.977	(0.882, 1.082)	0.948
Type of Enumeration Area	47.6			
Large MSA, Mailout/Mailback		1.000	reference level	1.000
Medium MSA, Mailout/Mailback		1.091	(1.001, 1.184)	1.248
Small MSA and Non-MSA MO/MB		1.010	(0.924, 1.104)	1.248
All other types of enumeration areas		0.735	(0.671, 0.804)	0.973
Interactions				
Large Household, not Reference Person	154.2	0.457	(0.404, 0.517)	N.A.
Black or Hispanic, Midwest	33.6	0.688	(0.606, 0.781)	N.A.
Multi-unit, West	18.8	1.417	(1.211, 1.660)	N.A.

- The variable **Relationship** stood out as an important variable, with reference people and their spouses more likely to be captured by the census than other household members.
- **Age/Sex**: older people and children were more likely to be captured than 18-49 year olds, and females were more likely to be captured than males.
- The **Size of Household**, along with its interaction with large household, was an important variable. People in large households had greater odds of capture, unless there were seven or more people, in which case people who were not the reference person had smaller odds of

capture. This effect was because the census form accommodated six people. Persons seven and higher were roster people who the A.C.E. counted as not captured by the census.

- **Tenure** played a bigger role in capture than it did in correct enumeration, with renters having only 0.605 the odds of capture as owners.
- **Type of Basic Address**: people in single units were more likely to be captured than people in multi-units or trailers.
- **Region**: people in the Midwest were more likely to be captured than those in other regions. However, this did not hold true for Blacks or Hispanics. Because of the interaction with Midwest the odds ratio was about one ($1.398 \times 0.744 = 1.04$).
- **TEA**: people in mailout/mailback areas were more likely to be captured. This contrasted with the univariate analysis, which found that people in Large MSA MO/MB areas had smaller odds of capture.
- The relationships between the variables Age/Sex, Tenure, Type of Basic Address, Racial/Ethnic Domain, and Region with odds of capture were shown by the P-sample model to be weaker than they appeared in the univariate analysis.

5. RECOMMENDATIONS

The analysis leads us to the following recommendations:

- The Executive Steering Committee for Accuracy Coverage and Evaluation Policy should view the results of these analyses as reassurance about the quality of the A.C.E. as our results were either as expected or interpretable.
- Future census researchers should use models such as logistic regression to gain insights into their multivariate data.

6. REFERENCES

Childers, Danny R. (2001): *Accuracy and Coverage Evaluation: The Design Document*. DSSD Census 2000 Dress Rehearsal Memorandum Series, Chapter S-DT-1.

Cromar, Ryan, and Farber, James (2000): *Accuracy and Coverage Evaluation: Final Large Block Cluster Subsampling Specifications.* DSSD Census 2000 Dress Rehearsal Memorandum Series, Chapter R-38.

Griffin, Richard, and Haines, Dawn (2000): *Accuracy and Coverage Evaluation Survey: Final Post-stratification Plan for Dual System Estimation.* DSSD Census 2000 Procedures and Operations Memorandum Series #Q-24.

Hosmer, David W., Lemeshow, Stanley (1989): *Applied Logistic Regression*. John Wiley and Sons, New York.

Ikeda, Michael (2000): *Accuracy and Coverage Evaluation Survey: Specifications for the Missing Data Procedures.* DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter Q-25.

Olson, Douglas (1999): *Accuracy and Coverage Evaluation Survey- Identification and Sampling of Block Clusters for Targeted Extended Search.* DSSD Census 2000 Dress Rehearsal Memorandum Series, Chapter R-38.

Wolter, Kirk M. (1986): *Some Coverage Error Models for Census Data*. Journal of the American Statistical Association, June 1986, Vol. 81, No. 394.

RTI (1997): *SUDAAN User's Manual, Release 7.5.* Research Triangle Institute, Research Triangle Park, North Carolina, 27709.

SAS (1994): *SAT/STAT User's Guide, Version 6, Fourth Edition Volume 2*. The SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

APPENDIX: Technical Documentation

Source Data Files

This section documents the source of the input data that was modeled.

E-Sample Model

E-Sample Person Estimation file:	all E-sample variables are obtained from this file except two as noted below
Sample Design File:	the Final Sampling Stratum
Hundred Percent Census Edited File (HCEF):	the Form Type and the Number of Units at the Basic Street Address

P-Sample Model

P-Sample Person Estimation file:	all P-sample variables were obtained from this file except two as noted below
P-Sample Housing Unit Estimation file:	the Number of Possible P-sample People in Household on Census Day
Sample Design File:	the Final Sampling Stratum

Variable Recoding

This section documents the recoding of variables from the source files. The variables Size of Household (for both the E-sample and P-sample models) were derived from other variables. The variables Units at Basic Street Address (UBSA) and Type of Basic Address were recoded from variables of the same name. All other variables were taken directly from the source files.

E-Sample Model

Units at Basic Street Address (UBSA)

If UBSA<=1 then UBSA = ‘Single Unit’;
Else if UBSA <=10 and UBSA > 1 then UBSA = ‘2 - 9 Units’;
Else UBSA = ‘10+ Units’;

Size of Household (SIZE)

If population count (INP) > 6 then SIZE = ‘Seven or More People’;
Else if INP < 7 and INP > 1 then SIZE = ‘Two to Six People’;
Else SIZE = ‘Single Person Household’;

APPENDIX: Technical Documentation

P-Sample Model

Type of Address (TOBA)
If TOBA = 'Housing Unit in a Special Place' or 'Other' then TOBA = 'Single Unit';
If TOBA = 'Multi-unit in a Special Place' then TOBA = 'Multi-unit';

Size of Household (SIZE)
If Number of Possible P-sample People in Household on Census Day (POSSPSC) > 6 then SIZE = 'Seven or More People';
If POSSPSC = 1 then SIZE = 'Single Person Household';
Else SIZE = 'Two to Six People';

Estimation of the Model

In this section we discuss technical details of estimating the logistic regression models. We used SUDAAN's (Shah, Barnwell and Bieler, 1997) Proc Logistic to estimate the logistic regression models. SUDAAN correctly estimated data from complex surveys such as the A.C.E., including calculating correct standard errors, confidence intervals and test statistics. We also used SAS (1989) software's Proc Logistic to estimate the models. SAS gave the correct maximum likelihood estimates, that is, the same estimates that SUDAAN yielded, though the standard errors, confidence intervals and test statistics SAS yielded were incorrect. However, SAS had more modeling capabilities than SUDAAN; for example, it generated useful statistics like the concordance rates. Also important, SAS was able to calculate estimates for models with large numbers of parameters, such as those with interaction terms. SUDAAN tended to run out of memory with such models and could not compute estimates. In SUDAAN we used the Taylor linearization method for variance estimation and the Wald chi-square statistic to test whether a variable was significant. We approached the modeling with backward selection. All variables turned out to be significant at the 0.1 alpha level in both the E-sample and P-sample modeling with the exception of Form Type. Both the P-sample and E-sample models reflected the weight after cluster sampling, large block subsampling (Cromar 2000) and Targeted Extended Search sampling (Olson 1999). The P-sample model also reflected the weight of the probability of residence and the non-interview adjustment (Ikeda 2000).

We also investigated first order interactions. Because of the large number of interaction terms (29 choose two or 406 for the census modeling, and 30 choose two or 450 for the A.C.E. modeling), and SUDAAN's unsuitability for estimating models with many parameters, we screened for important interaction terms with estimates generated by SAS. The stronger interaction terms were then estimated and tested for significance using SUDAAN.